Chapter 9

Watching You Watch Movies: Using Eye Tracking to Inform Cognitive Film Theory

Tim J. Smith

Psychology, Malet Street

Birkbeck, University of London

London, UK, WC1E 7HX

tj.smith@bbk.ac.uk

Phone: +44 (0)20 7631 6359

Fax: +44 (0)20 7631 6312

> *"The art of plastic composition consists in leading the spectator's attention through the exact path and with the exact sequence prescribed by the author of the composition. This applies to the eye's movement over the surface of a canvas if the composition is expressed in painting, or over the surface of the screen if we are dealing with a film-frame."*
>
> *(Eisenstein, 1948; pg. 148)*

One of the key intentions of cognitive film theory is to understand the cognitive processes involved in the viewing of a film and how this relates to the intentions and decisions of a filmmaker. Film theorists have applied an array of intellectual approaches to this question including the psychoanalytical, philosophical, close formal analyses of films and the application of cultural and social theories (Bordwell & Carroll, 1996; pg. 444). These theoretical explorations have generated many rich and detailed hypotheses about how filmmaker decisions may influence viewers but these hypotheses generally remain untested. In order to identify evidence in support of or in disagreement with these hypotheses, cognitive film theorists can appeal to the methods and theories of empirical psychology and the allied disciplines of cognitive science and cognitive neuroscience. In order for empirical psychology methods to be applied to questions of cognitive film theory a framework is required that demonstrates how questions in one discipline can be mapped into another.

In this chapter the Cognitive Computation Cinematics (CCC) approach will be presented. The CCC approach triangulates our understanding of how we watch films via three traditionally separate approaches: 1) cognitive psychology and associated methods of hypotheses testing; 2) computational methods in audiovisual analysis and computational modeling; and 3) the formal and statistical analysis of film (Cinematics; Salt, 2009). These methods can be combined to varying degrees depending on the question being investigated.

For example, in order to understand why certain types of cuts are more "invisible" to viewers than others a CCC approach may be to first run an empirical study: viewers could be instructed to detect cuts during a movie, reaction times measured and compared across different types of cuts (e.g. Smith & Henderson, 2008). Differences between detection rates could then be investigated by recording viewer eye movements across each cut and examining how primitive audiovisual features, such as motion and loudness can account for eye movements across the cut using computational methods (see Mital, Smith, Hill, & Henderson, 2011). Finally, the evolution of such cuts over time can be examined by identifying and statistically analyzing the prevalence of such cuts within a corpus of films (i.e. performing *cinemetrical* analysis; see Cutting, DeLong, & Nothelfer, 2010). By combining these three approaches either within a single project or collating results from different studies, the cognitive motivation for cinematic techniques and their history and function within film style can be identified.

To begin investigating questions of film viewing the empirical methods chosen must capture the dynamic interplay between filmmaker and viewer that is critical in the creation of the cinematic experience. As movies are composed of a rapid series of shots joined together by *cuts* (instantaneous transitions from one shot to another) with each shot lasting less than 4 seconds on average (Cutting, Brunick, DeLong, Iricinschi, & Candan, 2011) the methods used to gain insight into how directorial decisions influence viewer cognition must operate at the temporal resolution of seconds or milliseconds. Cognitive psychology and neuroscience offer various techniques that may be useful for probing viewer cognition: introspection/self-report, behavioral investigation (such as memory or reaction time tests), biophysiological recordings (e.g. heart rate monitoring or galvanic skin response), electrophysiology (e.g. event-related potentials - ERPs) and neuroimaging (see Smith, Levin, & Cutting, in press; for review). Experiments investigating comprehension of edited sequences typically present a

film clip and then test viewer memory after a delay of minutes (e.g. Frith & Robson, 1975). The relationship between comprehension and momentary perception of the edited sequence has to be inferred from the resulting memory rather than being tested on-line during the film experience. Other techniques such as functional magnetic resonance imaging (fMRI) provide a continuous measure of viewer brain activity during a film from which on-line cognitive processing can be inferred (e.g. Hasson et al., 2008). But the delay inherent in fMRI (the increase in blood-oxygen levels measured by fMRI takes about 2 to 3 seconds to register) makes it hard to attribute the influence of brief directorial decisions to a particular change in brain activity. What is needed is a real-time measure of how a viewer is watching and processing a film. Such a technique is provided by eye-tracking.

<1> What is Eye Tracking?

Eye tracking is the measurement of the movement of a viewer's eyes in relation to a visual array, whether the array is a real-world scene, a tabletop, stimuli presented on a computer or a cinema screen. Methods for recording the movements of the eye have been around for over a hundred years (Wade & Tatler, 2005) but up until the last twenty years these techniques have been highly invasive and uncomfortable for the user. For example, some of the earliest pioneering research into how we view natural scenes was performed by Russian scientist, Alfred Yarbus utilizing a technique involving a miniscule lens attached to the viewer's eye via a suction cup. The viewer's head was then clamped stationary and a beam of light shone onto the lens. As the eye moved the reflected beam traced a path on a photographic plate, recording the eye movements (Yarbus, 1967). A similarly invasive technique still used today involves inserting a coil of wire embedded in a contact lens onto a viewer's anaesthetized eye and placing their head in a magnetic field (*scleral contact lens/magnetic search coil*

tracking). Such techniques are highly precise but require the viewer's head to be stabilized and can be highly uncomfortable when used for prolonged periods.

Fortunately, modern video camera and computer technology has progressed to the point that an alternative non-invasive eye tracking technique is now available: video-based combined pupil/corneal reflection tracking. This technique exploits the fact that infra-red light shone on the human eye produces a very specific pattern of reflectance. Infra-red (IR) light is generated by warm light sources but is invisible to the human eye. If the human eye is illuminated by IR the light enters the pupil and is not reflected back, creating a dark pupil and refracts off the outside of the eye (the cornea) creating a glint known as corneal reflection. As the eye rotates the pupil moves with the eye but the glint always remains in the same position relative to the IR light source. By identifying the displacement of the pupil centre relative to the glint we are able to identify the precise vector of the eye's movement in two dimensions. These vectors can be *calibrated* relative to a 2D plane, such as a movie of computer screen by asking the participant to look at a series of points on the screen (typically 5 or 9). The computer uses these points to build a model of the eye's movements and to infer where the viewer's *point of regard* is, i.e. where their eyes are pointing, and project this back on to the screen as a gaze point[1].

<1> Why Record Viewer Eye Movements?

The physical experience of watching a film may outwardly appear to be a very passive activity. However, the viewer is highly active. The viewer must process the rapid sequence of audiovisual information, perceive what is represented on the screen, comprehend

[1] For further technical details of eye tracking technology and how to run an eye tracking study see Duchowsky (2007) or Holmqvist et al (Holmqvist et al., 2011).

the characters, spaces and actions depicted and engage in the construction of the narrative throughout the film. The only external evidence of this internal activity visible to an observer are facial expressions, bodily movements, physiological changes (e.g. heart rate, sweating, and pupil dilation), involuntary vocalizations (e.g. laughter, screams), and eye movements. A film viewer will move their eyes to different points on the screen about 2-5 times a second. During a typical 90 minute feature film this amounts to around 21,600 eye movements!

Every eye movement indicates a new phase of visual processing. Due to visual acuity limitations of the human eye we cannot see the whole visual scene in detail at the same time and must move our eyes to sequentially process the parts of the scene we are interested in. Unlike the photosensitive chip in a digital camera, the light sensitive surface at the back of the human eye (the retina) is not uniformly sensitive to the light projected on to it. The retina can only process high-resolution color information at its very centre due to the distribution of photoreceptive cells. There are two types of photoreceptor in the retina: rods, that are sensitive to light at low light levels; and cones, that are sensitive to color and light at normal light levels. The rods and cones are unevenly distributed across the retina with the periphery predominantly covered by rods and most of the cones concentrated in a small region at the centre of the retina called the *fovea.* This region only occupies about 2 degrees of a visual angle, roughly equivalent to the portion of the scene covered by a thumbnail held at arm's length. The resolution of the image we perceive is greatest when processed by cones so drops rapidly as the distance from the fovea increases. Five degrees away from the fovea resolution drops by 70% and by 20 degrees it has dropped by 90% (where 360° is a circle horizontally encircling the viewers head) (Wertheim, 1894). As a result, we only have access to high resolution color information about the part of the scene projected on to or close to the fovea.

When our eyes stabilize on a point in space (i.e. *fixate*), encoding of visual information occurs (Henderson & Hollingworth, 1999). Each fixation lasts on average 330ms

(when focused on a static visual scene; Rayner, 1998) and varies in duration with the complexity of visual stimuli and viewing task (Henderson, 2003). To process a new part of the scene the eyes must rotate so that the new target is projected on to the fovea. These rapid eye movements are known as *saccades* and have a duration of 20-50ms and cover a distance of about 4° (Rayner, 1998). If the target of our attention requires a large saccade (>30°) or is outside of our current field of view (120°) the eye movement will be accompanied by a head and/or body rotation (Tatler & Land, 2011). When sat in a movie theatre, the angle occupied by the screen is likely to be greater than 30° as the recommended minimum viewing angle for the back row of a cinema auditorium is 35° and the closer a viewer sits to the screen the larger the viewing angle from one side of the screen to the other will be (THX, 2012). For the majority of the audience members the screen will subtend a significantly larger viewing angle necessitating a head rotation along with saccadic eye movement to comfortably view the edges of the screen.

When we perform a saccadic eye movement our eyes are generally open but we do not perceive the world blurring across our retina. This is because our visual sensitivity effectively shuts down during a saccade via a process known as *saccadic suppression* (Matin, 1974). You can see the result of this process yourself by looking in a mirror as you saccade from one eye to another. If you watch somebody else do it you can see their eyes in flight but if you look at your own eyes you will only see the fixations.

The sequence of fixations and saccades creates a *scanpath*: a record of where the viewer's eyes were pointed during a viewing period, which elements of the scene they were most likely to have attended, perceived, encoded in memory and, also the parts of the scene that were not attended. Visual attention can *covertly* shift away from fixation in order to increase processing of peripheral features but studies in which participants are free to move their eyes have shown that covert attention does not exhibit such scanning behavior and is

instead only shifted to the target of the next saccade (Deubel & Schneider, 1996; Kowler, Anderson, Dosher, & Blaser, 1995). Processing of peripheral visual information is mostly reserved for selecting future saccade targets, tracking moving targets, and extracting *gist* about scene category, layout and vague object information (see Findlay & Gilchrist, 2003 for review). Therefore, a record of where a person has fixated will also be a good measure of what they have processed in detail (Henderson, 1992), an observation which is the cornerstone of all eye movement research.

Knowing a viewer's scanpath during a film sequence is important as the brief duration of most shots means that the viewer will only be able to attend to a small proportion of the screen area. In an average movie theatre with a 40 foot screen viewed at a distance of 35 feet, this region at the centre of our gaze will only cover about 0.19% of the total screen area. Given that the average shot length of most films produced today is less than 4 seconds (Cutting, Brunick, DeLong, et al., 2011), viewers will only be able to make at most 20 fixations covering only 3.8% of the screen area. This miniscule amount highlights how important it is for a filmmaker to know exactly where their audience is looking at every moment. If viewers fail to attend to the elements in a shot that convey the most important visual information the viewer will fail to understand the shot which may lead to increasing confusion and a lack of enjoyment.

<1> The Uncanny Ability to Know Where We Are Looking

The desire to know where viewers look during a film has been present in film theory for decades but the technological innovations necessary to allow eye tracking during film viewing have only recently occurred. The Russian film director and theorist, Sergei

Eisenstein wrote about visual perception and the intended eye movement patterns of his viewers in 1943 (Eisenstein, 1943). Eisenstein even included a hypothesized diagram of the path of viewer eye movements during a sequence from his film *Alexander Nevsky* (1938). The diagram shows how Eisenstein expected the viewers to follow key features of his compositions such as characters, actions, textures, and perspectives and how the rise and fall of the eye movements on the screen mirrors the movements in the soundtrack, creating "audiovisual correspondences" (Eisenstein, 1943; pp. 154-216). In her insightful 1980 article, Barbara Anderson (1980) reframed Eisenstein's gaze analysis in terms of a testable hypothesis about how the sequence should be viewed. However, at the time of writing eye tracking technology had only progressed as far as dealing with gaze on static images and she was left proclaiming "Such experimentation, carried out under controlled conditions, would not only add to our understanding of visual perception but would have exceedingly important implications in the area of film theory." (Anderson, 1980; pg. 26).

Similar prognoses of viewer gaze behavior litter editing theory. Edward Dmytryk described how to make a good cut to a point of view in terms of the time taken for viewers to shift their eyes: "To make the cut, then, we fix the frame in which the actor's eyes have 'frozen', add three or four frames more to give the viewer time to react and move his eyes as he follows the actor's look, at which point the cut is made." (Dmytryk, 1986; pg. 444). Three to four frames (125-167ms at 24 frames per second) is similar to the minimum time taken to perform a saccadic eye movement (100-130ms; Fischer & Ramsperger, 1984). Dmytryk had learnt the timing of his own saccades without ever seeing an eyetracker! This "uncanny facility to have your brain 'watch and note' your [own] eyes' automatic responses" (Pepperman, 2004; page 11) is thought to be one of the defining qualities of a good editor. Discussions of the predicted eye movements, their speed and limitations are common throughout editing theory (Block, 2001; Murch, 2001; Pepperman, 2004; Reisz & Millar,

1953). However, up until now, none of these intuitions or predictions about how viewers watch movies have been empirically tested.

<1> How Do People Watch Movies?

The earliest applications of eye tracking to film viewing were severely limited in the insight they could provide due to technical difficulties of recording accurate gaze behavior on a moving image and analyzing the resulting scanpaths. Some studies resorted to descriptions of individual gaze behavior for particular film sequences and made no attempt at quantifying differences (Klin, Jones, Schultz, Volkmar, & Cohen, 2002; Treuting, 2006). Such descriptions are referred to as *qualitative* and are distinguished from *quantitative* methods that aim to measure differences. An example of a qualitative analysis is presented by Treuting (2006). Treuting eyetracked 14 participants watching a range of clips from feature films including *Shawshank Redemption* (1994) and *Harry Potter and the Philosopher's Stone* (2001). She provided no quantification of their gaze behavior in relation to the different clips but described observable tendencies within clips such as the apparent prioritization of faces and moving objects, especially during a *Quidditch* match from the *Harry Potter* clip. Such descriptions of viewing behavior for particular clips are a useful starting point but there are so many factors contributing to the composition of each shot and its place within the narrative that might be driving viewer attention that it is hard to extrapolate viewer behavior to other clips. In order to do this we need to quantify the behavior of multiple viewers in relation to particular shot content or cinematic feature.

One way in which gaze behavior can be quantified is to measure the collective behavior of all viewers. This technique has proved surprisingly insightful as, unlike gaze behavior during static scene viewing the gaze behavior of multiple film viewers exhibits a remarkable degree of coordination (Dorr, Martinetz, Gegenfurtner, & Barth, 2010; Goldstein,

Woods, & Peli, 2006; Mital, et al., 2011; T. J. Smith & J. M. Henderson, 2008; Stelmach, Tam, & Hearty, 1991; Tosi, Mecacci, & Pasquali, 1997). Typically, in static visual scenes there is agreement in which parts of an image are of universal interest (e.g. faces, and task-relevant objects) but people do not look in these locations at the same time (S. K. Mannan, K. H. Ruddock, & D. S. Wooding, 1995). However, when watching a movie the gaze of multiple viewers exhibits *attentional synchrony*: the spontaneous clustering of gaze (T. J. Smith & J. M. Henderson, 2008). Figure 9.1 demonstrates a clear example of attentional synchrony during the teaser trailer for Dreamwork Animation's *Puss in Boots* (2011). Notice how the gaze points (red circles) occupy only a small portion of the screen at any one moment and the heatmap representing how densely clustered the gaze is mostly grouped into a single patch.

+++ Insert Figure 9.1 about here +++

Stelmach and colleagues were the first to observe attentional synchrony during film viewing (Stelmach, et al., 1991)[2]. They were interested in whether the gaze behavior of viewers could be used to decrease the bandwidth needed for video compression by predicting the areas of the screen most likely to receive fixation and only render those areas in detail. When they asked 24 participants to free-view 15 forty-five second video clips they observed a substantial degree of agreement amongst viewers in terms of where they looked. Goldstein, Woods and Peli (2007) showed 20 adults six long clips from Hollywood movies and found that for more than half of the viewing time the distribution of fixations from all viewers occupied less than 12% of the screen area. Attentional synchrony has subsequently been observed in a variety of moving-image types including feature films (Carmi & Itti, 2006b; Goldstein, et al., 2007; Hasson, et al., 2008; Marchant, Raybould, Renshaw, & Stevens, 2009; May, Dean, & Barnard, 2003; Nyström & Holmqvist, 2010; Smith, 2006; Smith &

[2] Although they referred to it as "a substantial degree of agreement among viewers in terms of where they looked" and not attentional synchrony (T. J. Smith & J. M. Henderson, 2008).

Henderson, 2008; Stelmach, et al., 1991; Tosi, et al., 1997), television (Sawahata et al., 2008), and unedited videos of real-world scenes (Cristino & Baddeley, 2009; Smith & Henderson, 2008; t' Hart et al., 2009).

In a systematic comparison of attentional synchrony across a variety of moving-image types, Dorr and colleagues demonstrated how the degree of attentional synchrony observed during Hollywood movies decreased during unedited videos of dynamic scenes (Dorr, et al., 2010). All types of moving-image contained moments when up to 80% of viewers looked at the same region of the screen at the same time but the proportion of overall viewing time during which this happened was significantly greater for professionally composed Hollywood movies than naturalistic videos. These findings suggest that composition and editing of movies causes attentional synchrony. To take this prediction forward we need to understand how visual features may influence where we fixate in a dynamic scene.

<1> Mise en Seen

Where we fixate in a visual scene is a consequence of the interaction between our internal plans, desires, and viewing tasks (known as *endogenous* control as it originates internally) and features of the audiovisual scene such as luminance, color, edges and motion (known as *exogenous* control as it originates externally) (Pashler, 1998)[3]. In cinema, the

[3] Audio features of a scene can also influence visual attention and where we fixate. In film, the inclusion of diegetic sounds, dialogue, off-screen sounds, and non-diegetic sounds such as soundtrack or narration may influence how viewers attend to the film. However, there is substantially less empirical research into the influence of audio features on visual attention and, as such I will mostly focus on visual influences for the current chapter. The combined influences of audiovisual features on viewing behaviour for moving-images is a research

exogenous factors can be thought of as a film's *mise en scène*: what appears in the film frame due to directorial decisions of setting, costume, lighting, and the staging of action (Bordwell & Thompson, 2001). Taken from the original French, mise en scène literally means "staging a scene" and is the result of the director's decisions about how the narrative action will be represented on screen and over the course of the movie. Individual decisions such as the color of a costume and its framing by the camera will influence the final cinematic image as presented to the viewer and to which they will respond by moving their eyes and perceiving the content. This relationship between mise en scène and attention has been hypothesised by film theorists Bordwell and Thompson (2001) who stated that elements of a film's mise en scène may pull our attention and our eyes to certain parts of the screen. In discussing which features are most pronounced they make reference to vision science:

> *"Most basically, our visual system is attuned to perceiving change, both in time and space. Our eyes and brains are better suited for noticing differences than for concentrating on uniform, prolonged stimuli. Thus aspects of mise-en-scene will attract our attention by means of changes in light, shape, movement, and other aspects of the image"* (Bordwell & Thompson, 2001; pg. 189)

This influence of basic visual features via a film's mise en scène has also been noted by Bruce Block (2001). He believes that viewers' eyes will be primarily attracted by movement, then by bright points on the screen and by faces (Block, 2001; pg. 132). In a qualitative attempt to test Block's hypotheses, Treuting (2006) looked for instances of motion, brightness and faces in her eye movement data. She confirmed the bias of gaze

---

topic ripe for future investigation (see Chapter 1). I great introduction to the role audio plays in the cinematic experience is Michel Chion's *Audio-Vision* (1990).

towards faces and also identified moments during the films when gaze appeared to be attracted by motion. However, she observed less evidence of the influence of brightness and color. Treuting's attempt at observing a relationship between visual features and gaze is a fine demonstration of the limitation of a qualitative approach. Without quantifying the independent influence of these visual features it is impossible to know whether they will replicate across movies. For example, Steven Spielberg's *Schindler's List* (1993) uses black and white cinematography to tell the story of one man's attempts to save Jews from the concentration camps in Nazi occupied Poland. In one striking scene, Spielberg picks out a small girl in the chaos of Kraków's ghetto by depicting her red coat in full color against the monochrome background. In a later scene we catch a glimpse of the girl's coat amongst a pile of dead bodies. A poignant use of color to single out a character amongst the faceless mass of lost souls. The contrast of the coat against the grey background is striking and probably results in quicker gaze to the girl than would have occurred without the red coat. However, the strength of the red coat as an attentional cue can only be known by quantifying the relative difference between it and the color of the entire image. Does the same use of a red coat on the daughter/mysterious figure in *Don't Look Now* (1973) result in similar capture of gaze even though the film is shot in color? The red coat seems to serve a similar purpose as in *Schindler's List* by guiding the viewer's eye to the small figure in the confusing alleyways and canals of Venice as Donald Sutherland's grieving father pursues the ghost of his recently deceased daughter. But the color red also serves a symbolic purpose in *Don't Look Now,* with red signifying the horror, death and mental anguish experienced by Donald Sutherland's character and his wife and the slow descent to tragedy that his pursuit of the red-clad figure represents. This symbolic function of red may outweigh its function as an attentional guide.

In order to progress from qualitative descriptions of how a visual feature such as the color red may influence viewer gaze to testable hypotheses we need to quantify the

relationship between visual features and viewer gaze behavior. Fortunately, computer vision provides us with tools to decompose any digital image into its constituent visual features such as brightness, color, edges, etc, and quantify their relationship to fixation location. Any digital color image, whether static or dynamic is stored as an array of pixels each with three or four-component color channels, RGB (Red, Green, Blue) or CMYK (Cyan, Magenta, Yellow, and Black). Each color channel has a value (typically from 0 to 255 for 8 bit color) specifying the amount of that color present in the pixel. The brightness (or *luminance*) of a pixel is created by the combination of the color channels and can be thought of as roughly equivalent to the amount of white in a grayscale version of the same image. The luminance and color channels are approximate to the light sensitivities of the photoreceptors in the human retina (Palmer, 1999). By combining these basic features in space and time computational algorithms can be used to identify low-level visual features such as oriented edges, corners, change over time ("flicker") or motion (Marr, 1982). Human early visual brain areas process a visual scene in a similar way and competition between these low-level features is believed to influence how we distribute our attention (Koch & Ullman, 1985). The weighted combination of low-level features is believed to create a *saliency map*: a viewpoint-dependent spatial map of the scene with a value at every location denoting how much that location "pops-out" and is likely to capture out attention exogenously (Itti & Koch, 2001). The highest points on this saliency map are selected as the target of the next saccade, the eyes move to that location and the saliency map is recomputed given the new viewpoint.

Initial evaluations of whether computational saliency maps could predict fixation location in static scenes showed some success. When participants look at still images without a viewing task low-level image properties such as edges, luminance contrast, and corners were significantly greater at fixation compared to control locations (Baddeley & Tatler, 2006; Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000; S. Mannan, K. H. Ruddock, & D. S.

Wooding, 1995; Mannan, Ruddock, & Wooding, 1996, 1997; Parkhurst & Niebur, 2003; Reinagel & Zador, 1999; Tatler, Baddeley, & Gilchrist, 2005). However, subsequent experiments have shown that static saliency does not influence fixation location when it is in conflict with the viewing task or scene semantics (Buswell, 1935; Castelhano, Mack, & Henderson, 2009; Einhauser, Spain, & Perona, 2008; Henderson, Brockmole, Castelhano, & Mack, 2007; Henderson, Malcolm, & Schandl, 2009; Torralba, Oliva, Castelhano, & Henderson, 2006; Yarbus, 1967). In a series of studies we have shown that even if the saliency of an object within a static scene is artificially increased by increasing its luminance (Walther & Koch, 2006) this does not increase the rate or time at which it is fixated if the viewer is looking for a different object (Smith & Henderson, 2010). Removing natural object salience also has no effect on fixation probability or time but boosting the salience of a looked-for object will guide the eyes to it (Smith & Henderson, 2010). This evidence suggests that the kinds of static low-level visual features contributing to a film's mise en scène have very little influence on gaze unless the salient objects are also of interest to the viewer, such as the girl with the red coat in *Schindler's List* (1993).

<1> Matching Action

Analysis of the influence of static low-level visual features on viewer gaze neglects the critical component distinguishing movies from photography: change over time. As noted by Bordwell & Thompson "*our visual system is attuned to perceiving change, both in time and space"* (Bordwell & Thompson, 2001; pg. 189). Change in space creates static feature contrast such as bright areas in a dark room, bold colors against a muted background or imbalance in the composition due to dense collection of edges (e.g. objects or texture) in one part of the frame. By comparison, change in time creates luminance or color changes and, most importantly, motion. Motion can either take the form of an optic flow field (Gibson,

1979) created by movement of the camera, movement of an object relative to the camera or a combination of the two. In film, the utility of motion for influencing viewer attention is widely known:

> *"Excluding cuts made at the beginnings and ends of sequences and self-contained scenes, cuts to reactions or responses, and cuts involving exchanges of dialogue, the cutter should look for some movement by the actor who holds the viewer's attention, and he should use that movement to trigger his cut from one scene to another. A broad action, will offer the easier cut, but even a slight movement of some part of the player's body can serve to initiate a cut which will be "smooth", or invisible…. The important consideration here is that there be just enough movement to catch the viewer's attention."* (Dmytryk, 1986, page 435-436)

What Dmytryk is describing is a technique of editing known as *match-action* (or *match-on-action*). A match-action cut is believed to be the smoothest way to transition between two viewpoints of an action and create *continuity* for the viewer: "the spectator's illusion of seeing a continuous piece of action is not interrupted" (Reisz & Millar, 1953, p. 216). Even the smallest movement, such as a head turn, shift of an actor's gaze, facial expression or eye blink are thought to offer an opportunity for a match-action cut (Murch, 2001; Pepperman, 2004; Reisz & Millar, 1953). In order to test the effectiveness of match-action edits we conducted an experiment in which participants were instructed to detect cuts in a series of 5 minute clips from feature films including *Blade Runner* (1982) and *Dogville* (2003). As predicted by match-action editing theory, participants failed to detect a third of all match-action cuts compared to only a tenth of between-scene cuts (Smith & Henderson, 2008). The sudden onset of motion before the match-action cut and the continuation of motion after the cut seems to mask the cut making it harder for viewers to detect it.

The relationship between such audiovisual events and the perception of cinematic continuity has been formalized in the *Attentional Theory of Cinematic Continuity* (AToCC; Smith, in press). AToCC argues that the critical component in the creation of continuity is viewer attention: the viewer needs to experience a clear flow of attention from the main content before the cut to the target of the next shot. The motivation for the cut needs to be established either through narrative, dialogue, off-screen audio cues, or motion and these cues guide viewer attention to the main content of the new shot. Motion plays a critical role in AToCC as it is assumed to drive attentional synchrony and provide a reliable cue that an editor can use to guide attention.

For example, in the teaser trailer for *Puss in Boots* (2011), the protagonist, Puss is depicted walking through a town and tossing first his hat to a bevy of appreciative female fans and then his sword to a group of children where it smashes open a piñata (Figure 9.1). The scene has a great sense of continuity of action but is constructed from six shots in quick succession. The impression of continuity is created by match-action editing: the first transition is a *whip pan* (rapid horizontal rotation of the camera) following the hat as it flies from Puss to the women with a cut to the women hidden within the pan. The second action is presented across two cuts, first to the sword in flight and the second to it hitting the piñata. In order to examine how this sequence of shots guides viewer attention and creates the apparent flow of continuity we eyetracked sixteen adults one at a time as they watched the trailer[4]. By superimposing the gaze of all sixteen participants back on to the movie and representing the

[4] Eye tracking was performed using an Eyelink 1000 desktop system (SR Research) with viewer's head stabilized on a chinrest at a viewing distance of 60 cm. The film was presented on a 21 inch screen at a screen resolution of 1280x1024 and a video resolution of 1280 x 720 at 24 fps (letter boxed). Heatmaps were created using CARPE (Mital, et al., 2011). Video of eye movements is available here: http://vimeo.com/25033301

density of their gaze on each frame of the trailer as a heatmap we can clearly see attentional synchrony within and across cuts (Figure 9.1). As Puss tosses his hat out screen left the camera pans to track it and viewer gaze saccades in the direction of the hat in an attempt to pursue it (Figure 9.1a,b,c,d). When the camera settles on the woman who catches the hat, the location of her head on the screen matches exactly where viewer gaze had been during pursuit, creating a smooth transference of attention from the hat to her face (e). Gaze then shifts back in the direction of travel fixating the other faces in the scene and finishing on the cat in the foreground: the true target of Puss's affections (f). Another whip pan takes us back to Puss and gaze shifts quickly back to him at screen center (g). Puss tosses his sword out of screen right, cuing a match-action cut to a very brief close-up of the sword in flight (h; lasting less than a second). The movement is too quick and the sword too small for the viewers to saccade to so the close-up shot instead positions the sword in the same screen location as Puss to create continuity of attention. Viewer gaze shifts slightly to screen right as the sword leaves screen and lands on the face of one of the children after the match action cut (i). After the sword has passed through the viewer's line of sight, smashing the piñata and embedding itself in the tree, viewer gaze reverses the pattern observed in shot f and saccades back in the direction of travel, fixating the faces of the remaining children in the scene (j). Throughout this rapid sequence the editor has precisely anticipated how the viewers will watch the scene, what features they will be interested and where their eyes will shift to in order to create continuity of action via viewer attention. This is the key method for continuity creation put forward in AToCC (Smith, in press).

Filmmakers' confidence in the power of motion to capture attention is supported by evidence from basic attention literature. Experiments using sparse stimuli or search of simple visual arrays have shown that motion is one of the strongest factors influencing visual attention irrespective of viewing task (Wolfe & Horowitz, 2004). However, such experiments

only tell us how motion works in relative isolation. In a film, motion is only one feature of a complex audiovisual scene. How do we know that an actor's sudden head turn will capture viewer attention during a film?

In a recent study, the *Dynamic Images and Eye Movements* (DIEM) project my colleagues and I investigated the influence of low-level visual features such as luminance, color, edges, and motion on gaze and attentional synchrony during moving-image viewing. We recorded eye movements of 251 people while they watched short high-definition TV and film clips taken from a broad range of categories including movie trailers, music videos, news, sports, documentaries, natural history, and educational videos[5]. The large range of clips and large number of viewers allowed us to examine where people looked during film viewing, how similar the gaze of multiple viewers was and which visual features predicted their gaze. All gaze data, source video clips, and resulting visualization of viewing behavior for each clip has been made publically available as part of an open-source corpus[6].

In order to understand what caused attentional synchrony we decomposed each frame of video into its constituent low-level visual features (luminance and two-opponent color channels) and then used common algorithms from computer vision to compute neurologically-plausible mid-level visual features such as edges, oriented edges, corners and

[5] Eye movements are recorded using an Eyelink 1000 (SR Research) eyetracker and videos presented on a 21'' Viewsonic Monitor with desktop resolution 1280x960@120Hz at a viewing distance of 90 cm. The location of the gaze can then be superimposed on to the associated frame to represent where the viewer was attending. See Mital et al (2011) for more details.

[6] For further information on the project checkout the DIEM website (http://thediemproject.wordpress.com/) and visualisations of gaze behaviour (http://vimeo.com/visualcognition/videos).

motion. The influence of each feature on viewer gaze was then calculated by comparing feature values around the locations of the video fixated by viewers to control locations. This analysis suggested that low-level features such as luminance and color are not predictive of gaze. By comparison, motion is highly predictive of gaze especially when a frame contains a single point of high motion against a static background (creating *motion contrast*). Such frames result in a high degree of attentional synchrony as all viewers look in the same place at the same time (Mital, et al., 2011). This suggests that movement in the scene is a strong factor in influencing where all viewers looked while free-viewing the videos in the DIEM corpus. A similar influence of motion and dynamic salience (a combination of motion and other static visual features) on gaze behavior during free-viewing has been shown by other studies (Berg, Boehnke, Marino, Munoz, & Itti, 2009; Carmi & Itti, 2006a, 2006b; Itti, 2005, 2006; Le Meur, Le Callet, & Barba, 2007; t' Hart, et al., 2009; Vig, Dorr, & Barth, 2009).

It is important to note that the critical component predicting gaze behavior is not motion per se, but motion contrast: how the distribution of motion varies across the frame. If every pixel on the screen has a high motion value due to camera movement, motion would not be predictive of gaze location. But when a small area of the screen is moving relative to a static background, the high degree of motion contrast is highly predictive of gaze location across all viewers, leading to attentional synchrony (Mital, et al., 2011). This evidence seems to support filmmaker intuitions about the power of small movements for attracting viewer attention and hiding a cut (Dmytryk, 1986; Murch, 2001; Pepperman, 2004; Reisz & Millar, 1953).

<1> Cut to the Chase

Exogenous influences on gaze are greatest immediately following a cut and decrease over the course of a shot as viewers become familiar with shot content (Carmi & Itti, 2006b;

Dorr, et al., 2010; Mital, et al., 2011). At the onset of a new shot saccade frequency (Germeys & d'Ydewalle, 2007; May, et al., 2003; Smith & Henderson, 2008) and attentional synchrony are highest and decrease over time (Carmi & Itti, 2006b; Dorr, et al., 2010; Mital, et al., 2011; Smith & Henderson, 2008). The decrease in exogenous control during a shot was predicted by Hochberg and Brooks (1978): "*visual momentum* is the impetus to obtain information and....should be reflected by the frequency with which glances are made... Visual momentum should presumably, decline with the length of time the viewer has been looking at the display, and should increase with the number of different places at which he can look to receive nonredundant information" (pg. 295). At the time, eyetracking technology did not permit Hochberg and Brooks to eyetrack viewers watching movies so instead they presented slideshows made up of images from old magazines, appliance catalogs and college yearbooks. They observed the predicted peak in saccade frequency at the onset of each image followed by a linear decrease until around 4 seconds and then the saccade frequency stopped decreasing (*asymptoted*) and remained low for the remainder of the time the images remained on the screen. The saccade frequency was higher for images with multiple centers of interest (mostly faces), for centers of interest offset from the screen center and for shorter presentation durations (Hochberg & Brooks, 1978). They believed the decrease in saccade frequency was direct evidence that each shot had limited information relevant to the viewer and after they had fixated all sources of information the shot became "cinematically dead" (pg. 294). An editor can optimize the visual momentum of a film by cutting to new information or reframing the old information once the viewer has exhausted the information. In this way, the editor can keep the image "alive", the viewer's gaze active and attentional synchrony at its highest throughout the film.

A similar change in saccade frequency over viewing time has been shown during static scene viewing (Antes, 1974; Buswell, 1935; Unema, Pannasch, Joos, & Velichovsky,

2005). Viewers are thought to initially go through an *ambient* phase of processing during which they perform large and frequent saccades around the image to construct an initial representation of the scene layout and content (Unema, et al., 2005). Over time, the viewer enters a *focal* phase of processing as the frequency of saccades and their amplitudes decrease and they spend longer fixating a small number of objects. Given that the scene is static, eventually the viewer will exhaust the informative content and cycle back to previously viewed areas (Yarbus, 1967). The rate at which a viewer shifts their gaze and the time they spend in each fixation is believed to be a factor of the information processed at fixation and the remaining information in the scene (see Nuthmann, Smith, Engbert, & Henderson, 2010; for a review of the factors influencing saccade timing in natural scenes).

By comparison, moving-images have the potential to constantly update the information in an image by moving the camera or the image content. However, increasing familiarity with the shot and decreasing impact of exogenous factors seem to result in increased variability between viewer gaze locations and a decrease in saccade frequency (Carmi & Itti, 2006b; Dorr, et al., 2010; Mital, et al., 2011). It is currently unclear whether viewers go through similar ambient and focal phases of processing when viewing movies but the change in attentional synchrony over time suggests a change of viewing strategy. The peak in attentional synchrony occurs 533ms following a cut indicating that the main features of the new shot are located with the first or second saccade (Mital, et al., 2011). If the shot ends soon after then attentional synchrony will be guaranteed. As the duration of the shot increases so does the variance between viewers' gaze. However, this does not mean that an average shot length of 533ms is optimum as the viewer needs time to comprehend the content of the new shot and not just locate the main features. Rapidly edited sequences such as movie trailers show a high degree of attentional synchrony but gaze is mostly stationary,

focussed at the screen center as each cut presents the new content in exactly the same screen location as the previous shot (Dorr, et al., 2010; Mital, et al., 2011).

Maintaining visual momentum and an optimal rate of information flow during a movie may not simply be a matter of matching shot duration to the content of that shot. The natural vacillations of viewer attention must also be considered. Cutting and colleagues have shown that Hollywood editing patterns have been evolving over time towards a nested pattern of shot lengths that may mirror the natural fluctuations of human attention (Cutting, Brunick, & DeLong, 2011; Cutting, Brunick, DeLong, et al., 2011; Cutting, et al., 2010). By identifying every shot in a corpus of 160 films from 1935 to 2010 they were able to decompose each film into a sequence of shots with varying durations. Patterns within this sequence were then identified by correlating the duration of each shot with the next shot (lag 1), the shot after that (lag 2), the shot after that (lag 3), until the end of each film (lag N). These correlations revealed an increasing tendency towards local clusters of similar shot durations in recent films. For example, high-energy action sequences tend to contain lots of short duration shots but are bracketed by shots of increasing duration as you move away from the period of high action. Similar patterns have been observed in human reaction time tests and are thought to govern the availability of attention for processing sensory information (Gilden, 2001). If Cutting and colleagues are right this will suggest that film is evolving to be compatible with the cognition of its viewers. Viewer attention may be the critical factor in ensuring the successful transmission of the audiovisual information of a movie into the mental experience of the viewer. Optimal communication may be accomplished by matching the rate of presentation of information to the spatiotemporal constraints of viewer attention both covertly, in terms of processing resources, and overtly in terms of where gaze is directed.

<1> Gazing at the Centre

As previously mentioned, rapidly edited sequences, such as movie trailers result in a large bias of viewer gaze towards screen center (Dorr, et al., 2010; Le Meur, et al., 2007; Mital, et al., 2011). This central tendency does not only occur for rapid sequences. A similar bias has been observed in static scenes (Tatler, et al., 2005) and is believed to be somewhat independent of image composition (Tatler, 2007). In movies, this centre bias is highly pronounced (Dorr, et al., 2010; Goldstein, et al., 2006; Le Meur, et al., 2007). In the DIEM project we described this centre bias and found that it seemed to be common across all videos irrespective of content, editing, or composition (Mital, et al., 2011). The center bias can clearly be seen in the distribution of gaze for a selection of videos from the DIEM corpus (

Figure 9.2; left column). Only when a video is composed with multiple centers of interest, such as the two boys playing in video 1 or the multi-frame composition of video 2 does the distribution of gaze shift away from the screen center. Otherwise the center bias is present in all videos, especially immediately following a cut, with the first saccade or two following a cut being biased towards screen centre. This bias decreases over the next second of a shot as viewers look at different parts of the frame. The center bias immediately following cuts also results in a high degree of attentional synchrony at screen center. In the DIEM corpus this has been expressed as *weighted cluster covariance*: optimal clusters describing the distribution of gaze for a particular frame are calculated and the size (*covariance*) and number of viewers in each cluster (*weight*) combined to make a single measure of attentional synchrony with lower values indicating more attentional synchrony and higher values less attentional synchrony (i.e. gaze is more spread out across the frame) (Mital, et al., 2011).

Figure 9.2 (right column) shows the frequency of different degrees of weighted cluster covariance for a particular movie. Films that have a greater center bias, such as the trailer for *Quantum of Solace* (2008;

Figure 9.2, film 4) or fewer focal objects (i.e. objects of interest), such as the two tennis players in film 3, show lower weighted cluster covariance due to a high degree of attentional synchrony.

The initial centre bias following a cut could simply be due to a tendency to frame objects of interest, such as faces at or near to screen centre. A systematic analysis of the factors contributing to the central bias in dynamic scenes confirmed that it is partly due to a bias in positioning focal, foreground objects at screen centre (Tseng, Carmi, Cameron, Munoz, & Itti, 2009). However, this study also showed a tendency for viewers to reset their gaze to the screen centre immediately following a cut irrespective of content.

+++ Insert Figure 9.2 about here +++

The centre bias of gaze confirms the artistic, photographic and cinematographic belief that the centre of a frame is a privileged position. In his classic work on spatial composition, *The Power of the Center,* Rudolf Arnheim (1988) discussed the aesthetic pleasure created by composing an image with the focal object at the centre of the frame. He believed the center holds stability and balance of a composition and placing an object at the center attributes the greatest visual importance to it. Arnheim's observation has been supported by analyses of the positioning of human faces in classic portraits (Tyler, 1998). Tyler found that one of the two eyes was typically centered along the vertical midline of a painting when the face was forward facing. When the face was depicted in profile the single visible eye or mouth tended to be along the vertical midline.

Tyler's work confirms artists' belief in the power of the center and their adherence to the convention but it does not prove that such central compositions are the most aesthetically pleasing. A competing compositional rule, known as the *Rule of Thirds,* states that the most

aesthetically pleasing compositions frame the focal object at the intersection of horizontal and vertical lines dividing the frame into thirds. Imagine a screen divided into three equally sized columns and three equally sized rows. These columns and rows intersect in four places: top right, top left, bottom right and bottom left. Artist intuition has claimed for centuries that the most aesthetically pleasing location for an object to be framed is either with the object's centre at the top right or top left location. Empirical evidence in support of the rule of thirds comes from studies comparing viewer preference for original paintings or their mirror-reversals (e.g. Levy, 1976). Viewers show a preference for paintings with their significant content on the right of the frame and this preference may be a product of hemispheric specialization as left-handed viewers show the opposite preference (Levy, 1976).

The apparent conflict between the center bias and the rule of thirds has recently been investigated in an elegant psychophysics study of aesthetic preference (Palmer, Gardner, & Wickens, 2008). Across a series of studies, Palmer and colleagues asked participants identify which of two alternative simple images they preferred. The images only differed in the placement of the focal object within the frame. When the object had a clear principle direction, such as a person or animal and was facing forward (i.e. towards the viewer) the most pleasing position was at the screen center. However, this center bias disappeared if the object was presented in profile: a left facing object was preferred at screen-right and a right facing object was preferred at screen-left. Palmer and colleagues concluded that aesthetic preference for composition depends on the direction at which the focal object is facing with viewers preferring the object to face on to the screen. This factor explains how the centre bias and rule of thirds can be reconciled depending on the facing direction of the focal object.

No such systematic empirical investigation of the aesthetic influence of composition in film currently exists. However, *AToCC* (Smith, in press) argues for the use of off-center placement of actor faces as a way of cuing viewer attention to the expectant space where the

target of the next shot will appear. Covert attention is cued in the direction of an actor's gaze, easing the transition across shots and creating continuity. In an example from *Requiem for a Dream* (Aronofsky, 2000), an actor facing off-screen is shown to lead to slowed orienting across cuts and difficulty in locating the focal object in the new shot (Smith, in press). However, while the actor's gaze direction in most close-up shots is on to the screen with the centre of their head positioned slightly off-center the location of their eyes may be close to screen center. As the actor's eyes will be the principle target of viewer gaze this may explain the center bias for viewer gaze we observe in the DIEM corpus (Mital, et al., 2011). Further, detailed analysis of existing films and empirical manipulations in the vein of Palmer, Gardner, and Wickens (2008) are required to test this hypothesis in film.

<1> Watching People Watch People

Analysis of the DIEM gaze data reveals a bias towards basic compositional features such as the screen centre and low-level visual features such as motion but it fails to identify any influence of the kinds of content we usually consider when discussing film e.g. people, actions and narratives. It is entirely plausible that being able to predict where we look based on motion does not necessarily mean that motion causes attention to shift to these locations. Motion may simply coincide with the features we are actually interested in. For example, looking at

Figure 9.2 (middle column) it is clear that gaze is mostly clustered around people and their faces. In most film and TV, people and animals are the main points of interest around which the shot will be composed. Drama emerges from the emotions, expressions, and thoughts portrayed in the character's face (Hitchcock, 1995). The principle task of the filmmaker is *"the organization of these oval shapes within the rectangle of the screen"* (Hitchcock, 1995). Careful shaping of shot composition, lighting, and focal depth will alter the low-level visual features of a shot and bias attention towards a face but the face itself is also a strong attractor of attention (Birmingham, Bischof, & Kingstone, 2008; Castelhano, Wieth, & Henderson, 2007; Yarbus, 1967). The motion of a person's face and body provides a source of potential information either through their interaction with the environment or their speech. The motion may predict where we look but we may look there because we are interested in people and their actions not the motion itself.

To look for the influence of people and faces on gaze behavior in the DIEM corpus all shots in a subset of videos from the corpus (see Mital, et al., 2011; for details) were categorized in terms of their shot size. Shot size or *camera-subject distance* is a common measure used in film theory and cinematography to describe how much of a human figure is present in a shot (Salt, 2009). For example, the three shots depicted in Figure 9.3 (Right-Bottom) increase in shot-size as they progress from a view of a person's face (Close-Up), to their upper body (Close Medium Shot) and whole body (Long Shot). If the main point of interest in most shots is a human face, as would be predicted from previous studies of gaze behavior in static scenes (Birmingham, et al., 2008; Castelhano, et al., 2007; Yarbus, 1967), then shot-size relative to a human figure should have a direct consequence on where viewers attend and how clustered gaze is for a particular shot. This is exactly what we observed in the DIEM corpus. When the shot did not contain a person ("NA" in Figure 9.3; Right-Top) gaze cluster covariance was the highest (i.e. least attentional synchrony). As the shot size

decreased, the cluster covariance also decreased. The shot size with the most attentional synchrony was Close Medium Shot. Such shots typically depict a single actor, framed centrally or slightly off-centre in conversation either with the camera (e.g. for news broadcasting; as in Figure 9.3; Right-Bottom-Centre) or an off-screen character. The actor's face occupies only a small area of the screen and is small enough to be seen in its entirety in a single fixation. To the best of my knowledge there is no film theory that would predict this primacy of Close Medium Shot's for maximizing coordination between how viewers attend to a shot.

+++ Insert Figure 9.3 about here +++

Once the shot size decreases past a Close Medium Shot the face enlarges and occupies more of the screen, forcing the viewer to saccade around the face to take in all the details (e.g. left eye, right eye, nose, mouth, etc). As there is no longer a single point of interest, cluster covariance increases again. This is clear evidence of how the gaze behavior of viewers and attentional synchrony is tied to the moment-by-moment information content of a film. Low-level features such as motion may predict where viewers are looking but the reason why viewers look there may actually be due to coincidence between motion and the social features viewers are actually interested in.

<1> Why are We Watching?

So far film viewing has been discussed as if it is a purely reactive task: cuts present us with new audiovisual information to which we respond with our eyes based on low-level visual features such as motion and seek out objects of interest such as faces. By this account, film viewers could be seen as dumb automaton without any volition or agency. However, we

are highly motivated viewers. We watch because we want to follow the narrative, comprehend the actions of characters, feel the intended emotions, and above all, enjoy the film. These motivations should provide strong endogenous drive to seek out information relative to the narrative. But is there any evidence of endogenous gaze control during film viewing? Bordwell & Thompson (2001) seem to think that there is:

> "...looking is purposeful; what we look *at* is guided by our assumptions and expectations about what to look *for*. These in turn are based on previous experiences of artworks and of the real world. In viewing a film image, we make hypotheses on the basis of many factors." (pg. 189)

The strong influence of endogenous factors on gaze behavior during static scene viewing has been known since the very earliest eye tracking studies (Buswell, 1935; Yarbus, 1967). The most famous early evidence of endogenous gaze control was conducted by Russian psychologist, Alfred Yarbus. Yarbus recorded viewer eye movements while they looked at a painting by Ilya Repin, *The Unexpected Visitor* (1884-88) depicting a man in military dress entering a sparsely decorated room and being greeted by a startled family. When the viewers were simply instructed to free-view the painting they spent most time looking at faces, clothing, foreground objects such as furniture but spent very little time looking at the background, walls, or floors (Yarbus, 1967). However, the key contribution of Yarbus' study was what he did next. He instructed the viewers to look at the painting six more times under different viewing instructions. Each instruction radically changed where viewers looked. Gaze was directed to the objects relevant to the viewing task such as faces for judging ages, clothes for remembering the clothing and furniture and background details when trying to remember the location of objects. Yarbus' evidence clearly showed that viewing task could have a direct influence on where we looked when viewing static scenes.

In moving-images, the heightened exogenous control by transients such as motion may mean that endogenous control has less of an influence on gaze location. This hypothesis would seem to be supported by the high degree of attentional synchrony observed during the viewing of edited moving-images (Dorr, et al., 2010; Goldstein, et al., 2006; Mital, et al., 2011; Smith & Henderson, 2008; Stelmach, et al., 1991; Tosi, et al., 1997). If gaze was under endogenous control, the individual variability in which image features a viewer prioritized at a particular moment would decrease attentional synchrony. Analysis of free-viewing cannot allow exogenous and endogenous factors to be dissociated as what viewers are interested in may also be what is visually salient.

To dissociate endogenous from exogenous factors either the viewing condition or the mental state of the viewer must be manipulated. For example, as the presentation time a dynamic scene increases viewer comprehension of the scene content, expectations about future events, and familiarity with visual features decreases the influence of exogenous factors and leads to more variability in gaze (Carmi & Itti, 2006a, 2006b; Dorr, et al., 2010; Mital, et al., 2011). As a result, unedited moving-images have less average attentional synchrony than edited sequences (Dorr, et al., 2010).

Familiarity with scene content can also be created over repeated viewings. Dorr and colleagues showed a decrease in attentional synchrony over repeated viewings of unedited naturalistic videos and Hollywood movie trailers (Dorr, et al., 2010). However, this may be a short term effect as repetitions separated by a day returned attentional synchrony back to the initial level (Dorr, et al., 2010). This decrease in attentional synchrony over repeated viewings may suggest less attention to salient features and more attention to the background. Such a finding would support the anecdotal observation that it is possible to notice new details of a movie with repeated viewings. It may also explain why continuity errors are easier to detect on repeated viewings: during the initial viewing gaze is driven by salient

features and only once knowledge of the scenes is accumulated can these salient features be ignored and attention be allocated to background features. This pattern of continuity error detection has been confirmed by Jon Sandys, the author of *Movie Mistakes* (Sandys, 2005) and expert on identifying and cataloguing continuity errors. Sandy's stated (in private communication) that most errors are initially detected as a "feeling that something isn't quite right" and only by replaying the scene can the error be properly identified. As calculated earlier, due to visual acuity limitations and time taken to move our eyes we can only fixate about 3.8% of the screen area during an average shot. This leaves lots of screen space to be explored on repeated viewings.

Another way of dissociating endogenous from exogenous control is to change the viewing task (a la Yarbus). In a preliminary investigation, we manipulated viewing task whilst viewers looked at unedited videos of natural scenes shot from a static camera position (Smith & Mital, 2011). Participants either watched the videos without a task the videos or attempted to recognize the location depicted in the video. In order to identify the location the viewers had to concentrate their gaze on the static features such as buildings, streets, signposts, trees, etc and ignore people and traffic. Viewers exhibited a surprising ability to ignore the moving features that had previously predicted gaze location during free-viewing. Gaze actively avoided people, was no longer predicted by motion and attentional synchrony decreased almost to the levels observed in static scenes (Smith & Mital, 2011). Even more surprising was what happened after participants pressed the button to signify recognition of the location: gaze immediately returned to following the motion. These preliminary results suggest that exogenous control can be overridden by the right viewing task but that our default interest is in people and their actions.

The absence of editing and deliberate composition of the dynamic scenes used in this study may explain how exogenous factors could be so readily overcome by viewing task.

Existing feature films should be used to examine whether viewing task has a similar effect on viewing behavior during normal movie. Spanne attempted such a manipulation using clips from *Armageddon* (1998) and *Die Hard* (1988) (Spanne, 2006). She instructed participants to either free-view the film clips and, decide if they wanted to view the rest of the film or form an opinion of the women that appeared in each clip. The results showed a lot less influence of viewing task on gaze than we observed in unedited sequences (Smith & Mital, 2011). The presence of an explicit viewing task lead to a decrease in attentional synchrony but the influence of task appeared to vary across clips and specific shot content (Spanne, 2006). The examples Spanne gives, such as a close-up of Bruce Willis' face in *Armageddon* (1998) which leads to low attentional synchrony in all conditions suggest that exogenous influences may fluctuate during a film. Directorial decisions such as mise en scène, staging and editing may influence the prominence of exogenous factors and their probability of wrestling gaze away from a competing viewing task. However, studies in this area are preliminary at best and further task manipulations and analyses of visual features around fixation are required to tease apart the endogenous/exogenous factors during film viewing.

In closing, the most important endogenous factor that may influence how we watch movies must be acknowledged: narrative. While film theory explorations of narratives possibly outnumber all other aspects of film put together, cognitive explorations of how we perceive cinematic narratives are virtually non-existent. A handful of cognitive psychologists have examined how we perceive and remember visual narratives (Kraft, 1987; Kraft, Cantor, & Gottdiener, 1991; Magliano & Zacks, 2011; Zacks & Magliano, 2009; Zacks, Speer, Swallow, & Maley, 2010) but, to my knowledge nobody has looked at how narrative comprehension influences how we watch movies. Given what we know about endogenous influences on gaze and the evidence presented for the accumulation of information during a dynamic scene it is highly likely that successful narrative comprehension should be evident in

viewer gaze. For example, do we seek out a character quicker when we know they are the murderer in a Film Noir? Do we search a scene for a bomb that we saw being hidden in an earlier scene? Do we gaze longer at a character that we empathize with? Do we refuse to look at something we anticipate to be alarming or uncomfortable (think of the dentistry/torture scene from *Marathon Man,* 1976)? The successful comprehension of a cinematic narrative requires the viewer to engage in the acquisition, comprehension and retention of the relevant information. This should be evident in viewer gaze but as yet this has not been demonstrated.

<1> Conclusion

To an external observer a film viewer may appear highly passive. The intention of this chapter was to demonstrate how incredibly active the viewer is both in terms of how they shift their gaze around the screen and cognitively process the presented information. The construction of the narrative is a collaborative process that requires suitable presentation of the relevant audiovisual information by the filmmaker and active acquisition and encoding of that information by the viewer. Many directorial decisions such as mise en scène, editing, and staging of action influence how the visual information is presented and how it may influence where a viewer looks exogenously. By applying a Cognitive Computational Cinematics (CCC) approach to film cognition this chapter has endeavored to confirm filmmaker intuitions about the influence of motion, feature contrast and faces on viewer attention using a combination of eye tracking and computer vision analyses of video content. These analyses suggest an interesting interaction between viewer comprehension and visual features such as motion and scene semantics that may fluctuate throughout a film. Eye tracking has the potential to provide a real-time insight into viewer cognition. Eye tracking can be used either in isolation or, in the future in combination with neuroimaging/electrophysiological methods. The intuitive nature of gaze data provides an immediate way into a viewer's experience of a film without having to engage with the complex quantitative aspects of empirical psychology.

However, once gaze data is broken down to its constituent eye movements and related to low-level or semantic features of a scene the potential for insight become limitless. I hope to be watching people watching people watch movies for a long time to come.

Acknowledgements

Thanks to Parag K. Mital for comments on an earlier draft and his assistance in the extended analysis of the Dynamic Images and Eye Movement data (DIEM: http://thediemproject.wordpress.com/) and Rachel Sandercocks for gathering data. The DIEM project was funded by the Leverhulme Trust (Ref F/00-158/BZ) and carried out with Prof. John M. Henderson, Robin Hill and in collaboration with Antje Nuthmann and Melissa Võ.

References

Anderson, B. F. (1980). Eye Movement and Cinematic Perception. *Journal of the University Film Association, 32*(1 & 2), 23-26.

Antes, J. R. (1974). Time Course of Picture Viewing. *Journal of Experimental Psychology, 103*(1), 62-70.

Arnheim, R. (1988). *The Power of the Center*. Berkeley, CA, USA: University of California Press.

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research, 46*, 2824-2833.

Berg, D. J., Boehnke, S. E., Marino, R. A., Munoz, D. P., & Itti, L. (2009). Free viewing of dynamic stimuli by humans and monkeys. *Journal of Vision, 9*(5), 1-15.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2008). Gaze selection in complex social scenes. *Visual Cognition, 16*, 341-355.

Block, B. (2001). *The Visual Story: Seeing structure of film, TV, and New Media*. MA, USA: Focal Press.

Bordwell, D., & Carroll, N. (1996). *Post-theory: Reconstructing film studies*. Madison, Wisconsin, USA: University of Madison Press.

Bordwell, D., & Thompson, K. (2001). *Film Art: An Introduction* (Vol. 6th). New York, USA: Mc Graw Hill.

Buswell, G. T. (1935). *How people look at pictures : a study of the psychology of perception in art*. Chicago, IL: The University of Chicago Press.

Carmi, R., & Itti, L. (2006a). The role of memory in guiding attention during natural vision. *Journal of Vision, 6*, 898-914.

Carmi, R., & Itti, L. (2006b). Visual causes versus correlates of attention selection in dynamic scenes. *Vision Research, 46*(2006), 4333.

Castelhano, M. S., Mack, M., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision, 9*, 1-15.

Castelhano, M. S., Wieth, M. S., & Henderson, J. M. (2007). I see what you see: Eye movements in real-world scenes are affected by perceived direction of gaze. In L. P. a. E. Rome (Ed.), *Attention in Cognitive Systems* (pp. 252-262). Berlin: Springer.

Chion, M. (1990). *Audio-Vision: Sound on Screen*. New York, US: Columbia University Press.

Cristino, F., & Baddeley, R. (2009). The nature of the visual representations involved in eye movements when walking down the street. [Article]. *Visual Cognition, 17*(6-7), 880-903. doi: DOI: 10.1080/13506280902834696

Cutting, J. E., Brunick, K. L., & DeLong, J. E. (2011). The changing poetics of the dissolve in Hollywood film. *Empirical Studies in the Arts, 26*, 149-169.

Cutting, J. E., Brunick, K. L., DeLong, J. E., Iricinschi, C., & Candan, A. (2011). Quicker, faster, darker: Changes in Hollywood film over 75 years. *i-Perception, 2*, 569-576.

Cutting, J. E., DeLong, J. E., & Nothelfer, C. E. (2010). Attention and the evolution of Hollywood film. *Psychological Science, 21*(3), 440-447.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vision Research, 36*(12), 1827-1837.

Dmytryk, E. (1986). *On Filmmaking*. London, UK: Focal Press.

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision, 10*(28), 1-17.

Duchowsky, A. (2007). *Eye tracking methodology: theory and practice* (2nd ed.). London: Springer-Verlag.

Einhauser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8*(14), 11-26.

Eisenstein, S. M. (1943). *The Film Sense* (J. Leyda, Trans.). London, UK: Faber and Faber.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing.* Oxford: University Press.

Fischer, B., & Ramsperger, E. (1984). Human express saccades: extremely short reaction times of goal directed eye movements. *Experimental Brain Research, 57*, 191-195.

Frith, U., & Robson, J. E. (1975). Perceiving the language of films. *Perception, 4*(1), 97-103.

Germeys, F., & d'Ydewalle, G. (2007). The psychology of film: perceiving beyond the cut. *Psychological Research, 71*(4), 458-466.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, USA: Houghton Mifflin.

Gilden, D. L. (2001). Cognitive emission of 1/f noise. *Psychological Review*(108), 33-56.

Goldstein, R. B., Woods, R. L., & Peli, E. (2006). Where people look when watching movies: Do all viewers look at the same place? *Computers in Biology and Medicine, 37*(7), 957-964.

Goldstein, R. B., Woods, R. L., & Peli, E. (2007). Where people look when watching movies: Do all viewers look at the same place? [Article]. *Computers in Biology and Medicine, 37*(7), 957-964.

Hasson, U., Landesman, O., Knappmeyer, B., Valines, I., Rubin, N., & Heeger, D. J. (2008). Neurocinematics: The Neuroscience of Film. *Projections: The Journal of Movies and Mind, 2*(1), 1-26.

Henderson, J. M. (1992). Visual attention and eye movement control during reading and picture viewing. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and reading* (pp. 260-283). New York: Springer-Verlag.

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Sciences, 7*(11), 498-504.

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual search in real-world scenes. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 537-562). Oxford: Elsevier.

Henderson, J. M., & Hollingworth, A. (1999). The role of fixation position in detecting scene changes across saccades. *Psychological Science, 10*(5), 438-443.

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review, 16*, 850-856.

Hochberg, J., & Brooks, V. (1978). Film Cutting and Visual Momentum. In J. W. Senders, D. F. Fisher & R. A. Monty (Eds.), *Eye Movements and the Higher Psychological Functions* (pp. 293-317). Hillsdale, NJ: Lawrence Erlbaum.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye Tracking: A comprehensive guide to methods and measures*. Oxford, UK: OUP Press.

Itti, L. (2005). Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition, 12*(6), 1093-1123.

Itti, L. (2006). Quantitative modelling of perceptual salience at human eye position. *Visual Cognition, 14*(4-8), 959-984.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194-203.

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry, 59*(9), 809-816.

Koch, C., & Ullman, S. (1985). Shifts in Selective Visual-Attention - Towards the Underlying Neural Circuitry. *Human Neurobiology, 4*(4), 219-227.

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The Role of Attention in the Programming of Saccades. *Vision Research, 35*(13), 1897-1916.

Kraft, R. N. (1987). Rules and strategies of visual narratives. *Perceptual and Motor Skills, 64*(1), 3-14.

Kraft, R. N., Cantor, P., & Gottdiener, C. (1991). The Coherence of Visual Narratives. *Communication Research, 18*(5), 601-615.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: an investigation with higher-order statistics. *Spatial Vision, 13*(2-3), 201-214.

Le Meur, O., Le Callet, P., & Barba, D. (2007). Predicting visual fixations on video based on low-level visual features. *Vision Research, 47*, 2483-2498.

Levy, J. (1976). Lateral dominance and aesthetic preference. *Neuropsychologia 14*, 431-445.

Magliano, J. P., & Zacks, J. M. (2011). The Impact of Continuity Editing in Narrative Film on Event Segmentation. *Cognitive Science, 35*(8), 1--29.

Mannan, S., Ruddock, K. H., & Wooding, D. S. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. [Article]. *Spatial Vision, 9*(3), 363-386.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. *Spatial Vision, 9*, 363-386.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. [Article]. *Spatial Vision, 10*(3), 165-188.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997). Fixation sequences made during visual examination of briefly presented 2D images. [Article]. *Spatial Vision, 11*(2), 157-178.

Marchant, P., Raybould, D., Renshaw, T., & Stevens, R. (2009). Are you seeing what I'm seeing? An eye-tracking evaluation of dynamic scenes. *Digital Creativity, 20*(3), 153-163.

Marr, D. C. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.

Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin, 81*(12), 899-917.

May, J., Dean, M. P., & Barnard, P. J. (2003). Using film cutting techniques in interface design. *Human-Computer Interaction, 18*, 325-372.

Mital, P. K., Smith, T. J., Hill, R. L., & Henderson, J. M. (2011). Clustering of gaze during dynamic scene viewing is predicted by motion. *Cognitive Computation, 3*(1), 5-24.

Murch, W. (2001). *In The Blink Of An Eye: a perspective on film editing*. Los Angeles, USA: Silman-James Press.

Nuthmann, A., Smith, T. J., Engbert, R., & Henderson, J. M. (2010). CRISP: A computational model of fixation durations in scene viewing. *Psychological Review, 117*(2), 382-405.

Nyström, M., & Holmqvist, K. (2010). Effect of compressed offline foveated video on viewing behavior and subjective quality. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), 6*(1), 1-16.

Palmer, S. E. (1999). *Vision Science: Photons to Phenomenology*. Boston, Massachusetts: MIT Press.

Palmer, S. E., Gardner, J. S., & Wickens, T. D. (2008). Aesthetic issues in spatial composition: Effects of position and direction on framing single objects. *Spatial Vision, 21*(3-5), 421-449.

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision, 6*, 125-154.

Pashler, H. (1998). *Attention*. Hove, UK: Psychology Press Ltd.

Pepperman, R. D. (2004). *The Eye is Quicker: Film Editing: Making a good Film better*. Studio City, CA, US: Michael Wiese Productions.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372-422.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network: Computer and Neural Systems, 10*, 1-10.

Reisz, K., & Millar, G. (1953). *Technique of Film Editing*. London, UK: Focal Press.

Salt, B. (2009). *Film Style and Technology: History and Analysis* (Vol. 3rd). Totton, Hampshire, UK: Starword.

Sawahata, Y., Khosla, R., Komine, K., Hiruma, N., Itou, T., Watanabe, S., et al. (2008). Determining comprehension and quality of TV programs using eye-gaze tracking. *Pattern Recognition, 41*(5), 1610-1626.

Smith, T. J. (2006). *An Attentional Theory of Continuity Editing.* Ph.D., University of Edinburgh, Edinburgh, UK.

Smith, T. J. (in press). Attentional Theory of Cinematic Continuity. *Projections*.

Smith, T. J., & Henderson, J. M. (2008). Attentional synchrony in static and dynamic scenes. *Journal of Vision, 8*(6), 773.

Smith, T. J., & Henderson, J. M. (2008). Edit Blindness: The Relationship Between Attention and Global Change Blindness in Dynamic Scenes. *Journal of Eye Movement Research, 2*(2), 1-17.

Smith, T. J., & Henderson, J. M. (2010). *The causal influence of visual salience on gaze guidance during scene search and memorisation.* Paper presented at the Object, Perception, Attention and Memory, St. Louis, Missouri.

Smith, T. J., Levin, D., & Cutting, J. E. (in press). A Window on Reality: Perceiving Edited Moving Images. *Current Directions in Psychological Science*.

Smith, T. J., & Mital, P. K. (2011). Watching the world go by: Attentional prioritization of social motion during dynamic scene viewing. [conference abstract]. *Journal of Vision, 11*(11), 478.

Spanne, J. G. (2006). *Task impact on cognitive processing of narrative fiction film.* Masters, Lund University, Lund.

Stelmach, L. B., Tam, W. J., & Hearty, P. J. (1991). *Static and dynamic spatial resolution in image coding: an investigation of eye movements.* Paper presented at the Human Vision, Visual Processing, and Digital Display II.

t' Hart, B. M., Vockeroth, J., Schumann, F., Bartl, K., Schneider, E., König, P., et al. (2009). Gaze allocation in natural stimuli: Comparing free exploration to head-fixed viewing conditions. *Visual Cognition, 17*(6/7), 1132-1158.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14), 1-17.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. [Article]. *Vision Research, 45*(5), 643-659. doi: doi:10.1016/j.visres.2004.09.017

Tatler, B. W., & Land, M. F. (2011). Vision and the representation of the surroundings in spatial memory. *Philosophical Transactions of the Royal Society B, 366*(596-610).

THX. (2012). THX tech pages Retrieved 13th February, 2012, 2012, from http://www.cinemaequipmentsales.com/athx2.html

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*(4), 766-786.

Tosi, V., Mecacci, L., & Pasquali, E. (1997). Scanning eye movements made when viewing film: Preliminary observations *International Journal of Neuroscience, 92*(1/2), 47-52.

Treuting, J. (2006). Eye tracking and cinema: A study of film theory and visual perception. *Society of Motion Picture and Television Engineers, 115*(1), 31-40.

Tseng, P. H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying centre bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*(7), 1-16.

Tyler, C. W. (1998). Painters centre one eye in portraits. *Nature, 392*, 877-878.

Unema, P. J. A., Pannasch, S., Joos, M., & Velichovsky, B. M. (2005). Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition, 12*, 473-494.

Vig, E., Dorr, M., & Barth, E. (2009). Efficient visual coding and the predictability of eye movements on natural movies. *Spatial Vision, 22*(2), 397-408.

Wade, N. J., & Tatler, B. W. (2005). *The moving tablet of the eye: The origins of modern eye movement research*. New York: Plenum press.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*, 1395-1407.

Wertheim, T. (1894). Über die indirekte Sehschärfe. *Z Psychologie, Physiologie, Sinnesorg, 7*(1), 121-187.

Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5*, 1-7.

Yarbus, A. L. (1967). *Eye Movements and Vision*. New York: Plenum Press.

Zacks, J. M., & Magliano, J. P. (2009). Film understanding and cognitive neuroscience. In D. P. Melcher & F. Bacci (Eds.). New York:: Oxford University Press.

Zacks, J. M., Speer, N. K., Swallow, K. M., & Maley, C. J. (2010). The brain's cutting-room floor: segmentation of narrative cinema. *Frontiers in Human Neuroscience, 4*.

Filmography

Aronofsky, D. (2000). *Requiem for a Dream,* USA

Bay, M. (1998). *Armageddon*, USA

Columbus, C. (2001). *Harry Potter and the Philosopher's Stone*, USA

Darabont, F. (1994). *Shawshank Redemption*, USA

Eisenstein, S. (1938). *Alexander Nevsky,* Russia

Forster, M. (2008). *Quantum of Solace*, UK

McTiernan, J. (1988). *Die Hard,* USA

Miller, C. (2011) *Puss in Boots*, USA

Roeg, N. (1973). *Don't Look Now*, USA

Schlesinger, J. (1976). *Marathon Man,* USA

Scott, R. (1982). *Blade Runner*, USA

Spielberg, S. (1993). *Schindler's List*, USA

Von Trier, L. (2003). *Dogville*, Denmark

Figure 9.1: Gaze behavior of sixteen viewers during a clip from the teaser trailer for *Puss in Boots* (Dreamwork Animation, 2011). Gaze point for each viewer is represented as a small red circle. The heatmap represents the distribution of gaze for that frame: hotter colors indicate more fixations in that area of the screen.

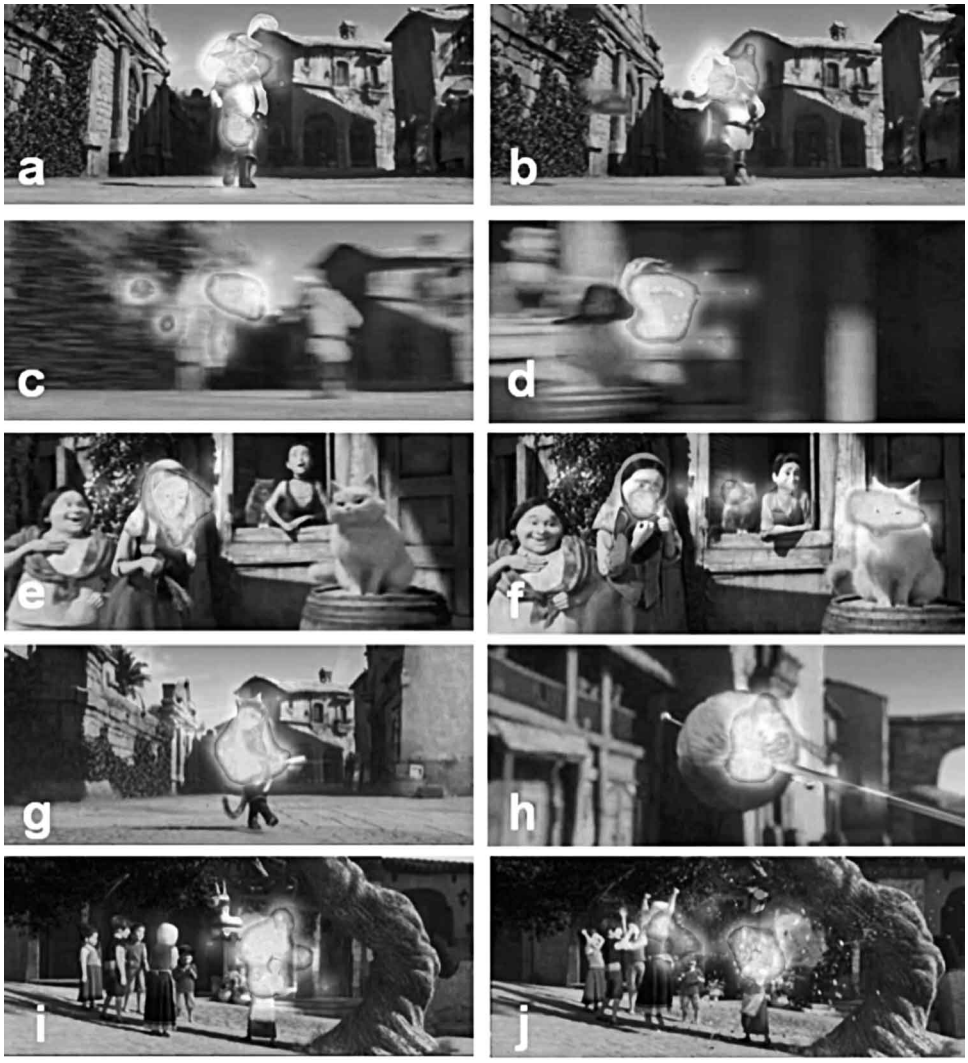

Figure 9.2: Sample videos taken from the DIEM database with superimposed gaze locations of 42 viewers for that particular frame (middle column). The clustering of the gaze is represented as a circular heatmap (middle column) and the covariance (i.e. spread) of the cluster is used to measure how much *attentional synchrony* is displayed across all the viewers. The distribution of cluster covariances allow us to see differences between videos (right column). The left column displays the distribution of gaze throughout the video as a heatmap: hotter colors indicate more fixations in that area of the screen. Figure modified with permission from Mital, Smith, Hill, & Henderson (2010).

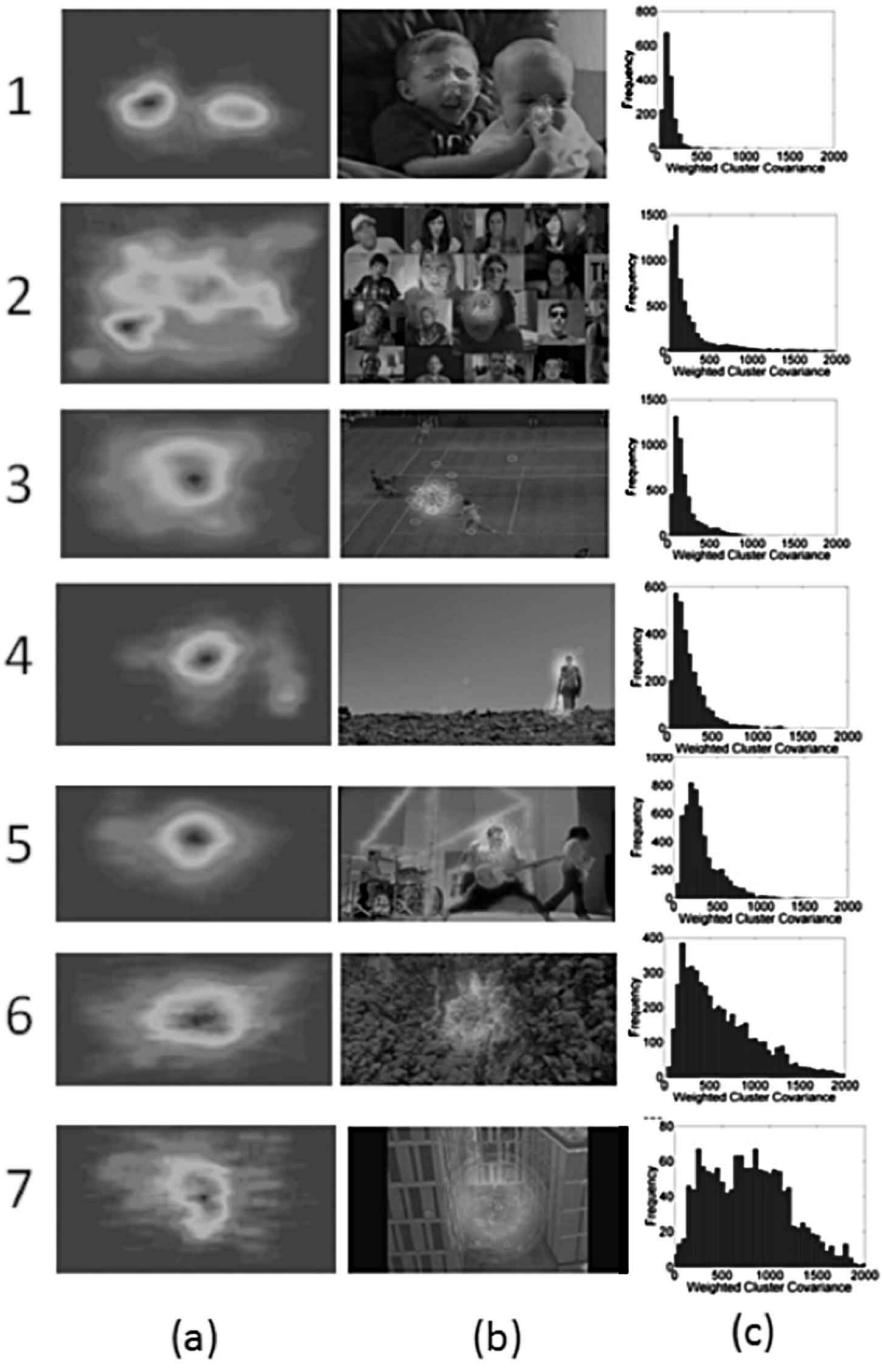

Figure 9.3: Right: Shot Size gauge. All shot sizes are specified relative to an average human figure as if they were positioned at the main focal depth of the shot and the top of the frame were aligned with the top of their head. Right-Top: gaze cluster covariance as a function of shot size. Right-Bottom: example frames of a Close-Up (CU), Close Medium Shot (CMS), and Long Shot (LS).